DOCUMENT RESUME

ED 223 668 TM 820 793

AUTHOR Milton, Ohmer
TITLE Will That Be on the Final?
PUB DATE Aug 82
NOTE 8p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 23-27, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; *Evaluation Methods; Higher Education; Learning Activities; *Teacher Made Tests; *Test Construction; Test Format; *Testing Problems; Test Items; Test Reliability; Test Validity

ABSTRACT

Educators are called upon to improve the quality of classroom tests to enhance the learning of content. Less faculty concern for tests than for other features of instruction, compounded by a lack of knowing how to assess different levels of learning with test questions that measure complex processes, appear to generate poor quality classroom tests. Research indicates that publisher supplied items and many instructors' tests include mostly recall items requiring students to memorize details and to study little else. This results in a deficiency of analytical skills (the higher order processes of application and evaluation) and a decrease in content learning. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223668

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

WILL THAT BE ON THE FINAL?

Ohmer Milton, Director
Learning Research Center and
Professor of Psychology
University of Tennessee, Knoxville

The essential thesis of my remarks is rather straightforward--classroom testing is a powerful instructional method; improved classroom tests will enhance the learning of content. Yet it is rather ironic that we as psychologists, who have been leaders in creating principles and procedures for the development of standardized tests for well over half a century and are learning specialists, have tended to ignore the influences of classroom tests upon the learning by our undergraduate charges. Just last year this Association issued a special number of the AMERICAN PSYCHOLOGIST entitled "Testing: Concepts, Policy, Practice, and Research" (Glaser and Bond, 1981). It contained 20 invited articles; not a single one dealt with classroom tests.

But, in conservative round numbers, with 10 million students in colleges and universities and each one taking three courses per term and having two tests in each course, some 60 million classroom tests are given every three or four months--or a total of almost 200 million in an academic year. Even though my arithmetic may be in error by several million, the numbers are nonetheless staggering. Now that we have personal computers along with the giant ones, just think of all the myriad statistical manipulations that await us! Although many facets of instruction have received research attention, tests as important teaching-learning endeavors have been slighted.

There is no way of even guessing about the numbers of these classroom test questions which come from the test banks or manuals accompanying textbooks. We can say that the numbers are huge--in our own field and in many others--because they are provided with most introductory texts and with a few advanced ones.

A few studies suggest that classroom tests in general and test bank ones in particular tend to be of poor quality (Milton, 1982). Because of the theme of this session, I'll focus primarily on studies of the latter.

McMillan (1979) surveyed faculties of marketing courses in colleges

---

Paper presented at the Annual Convention of the American Psychological Association, August, 1982, Washington, D. C.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

O. Milton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

and universities in 43 states and the District of Columbia. Most agreed that the major inadequacies of publisher supplied multiple-choice items included: 1) focusing on trivia, 2) some part of the item being stated ambiguously or phrased awkwardly, 3) two or more alternatives being correct, and 4) no answer being correct.

Evans, Dodson, and Bailey (1981) examined the manuals for six psychology tests. Every fifth multiple-choice item, from a total of almost 2,000, was drawn and then classified independently by two judges--there was almost complete agreement between them. Ninety-five percent of the items required students to recognize only simple factual information. This finding is quite consistent with those from a few studies of locally prepared course items. For example, a review of tests used in medical schools (Wingard and Williamson, 1973) revealed that most course exams require recognition or recall of isolated bits of information.

Mentzer (1982) inspected 35 sets or almost 26,000 psychology items for cues or tips to the correct answers--such cues as the longest alternative being the correct one and similarities between portions of the stem and the correct alternative. In 17 of the sets, one-fourth or more of them contained cues, and in an additional ten sets, one-third or more contained them.

Milton and Eison (1982) report finding a manual consisting of 366 multiple-choice items in which one-third of them contained both "All of the above" and "None of the above" as alternatives; an additional one-fifth contained either one or the other. In another manual of 900 items, almost one-third of them were stated negatively. In several of the chapters, as many as one-half the items were negatives and in two of them, two-thirds were.

Now just a very few examples of the sorts of poor quality items to which I've been alluding. This first one is not only an illustration, it also contains a bit of erudition that most psychologists probably do not know, but should! If you remember nothing else of what I say, I hope you remember the answer to this question:

Tradition has it that the Compromise of 1877 was worked out:

a. by President Ulysses S. Grant

b. between Hayes and Tilden

c. at the Wormley Hotel

d. both a. and c.

The psychology examples will be true-false ones; I'll read these one at a time and let's have a show of hands for the correct answer of each one:

T F "The amount of time spent clinging to mother" would be an operational definition of love.

T F Six-day-old infants are often capable of imitating the behavior of others.

T F All neurons are surrounded by a myelin sheath.

T F The polygraph is a completely accurate lie detector.

By my reckoning the first two items are ambiguous and the second two contain "absolutes" as cues--rendering them False.

I am inclined to think that most instructors show great concern for most aspects of instruction, but generally are unaware of the importance of tests to students. In addition to my own observations of faculty concerns, based upon talking with many of them over the years, here is a bit of evidence of a different sort, albeit crude.

Some of you who are familiar with questionnaires for student evaluation of faculty know that the Purdue Cafeteria System consists of a catalogue of 200 statements about classroom instruction. Each instructor chooses from those 200 around 35 that reflect those aspects about which he or she is most interested in having students make judgments. A representative of the Center for Instructional Services at Purdue (Cree, 1982) gave me the numbers of classrooms during the five-year period 1974-1979 for which selected statements had been chosen. These numbers do not refer to numbers of faculty because a given member may be represented more than once. There were 9,033 classrooms. Here are a few statements which were among those chosen most often by faculty for student evaluations and those which were chosen least often.

Most

	No. of classrooms choosing
My instructor displays a clear understanding of course topics.	6,673
My instructor displays enthusiasm when teaching.	6,327
My instructor seems well-prepared for class.	5,107
My instructor has an effective style of presentation.	4,850

Least

Exams are free from ambiguity.	238
Exams are creative and require thought.	430
Exams are coordinated with major course objectives.	819
Exams cover a reasonable amount of material.	902

I remind you again that these data are crude, but they certainly point in the direction of less faculty concern for course tests than for other features of instruction. Much more careful analyses and on a grander scale of the instructional judgments we seek from studemts might tell us more than we'd like to know about ourselves.

There may be still another ingredient among many faculty that helps to account for poor quality tests--that of not knowing how to assess different levels of learning with different sorts of test questions. In a carefully controlled study (Semb and Spencer, 1976) at a large university, faculty representing several fields were asked to indicate the percentages of their test items that measured complex processes-problem solving, generalizing, etc.--versus those tapping simple recall.

The instructors estimated that 31.5 percent of their items were complex, but independent trained judges found that only 8.5 percent of the items on major exams were complex. This finding of around 90 percent of course items calling for factual information is quite consistent with results in other studies (Milton, 1982). Semb and Spencer concluded: "In general, it would appear as if many instructors are testing mostly over recall tasks. What is disturbing is that they do not even know it." (P. 121.)

Now for evidence supporting the proposition that tests are powerful influences upon the learning of content. I commend to your attention a review article by Balch (1964); he concluded that the use of a test--as opposed to not using one--and the nature of the exam affected students' learning significantly. For example, if a test will consist mainly of questions requiring knowledge of simple factual information, students will memorize details and study little else. I also call to your attention the most recent study (Halpin and Halpin, 1980) I could find along these lines. The study is a model one in dealing with the variables. At the same time, an additional dependent variable was investigated--the influence of tests upon student ratings of instruction; course achievement goes up and ratings of instruction go down.

A deficiency of the "test influence" research is the omission of the higher order processes of application and evaluation, for example. In this connection, the work of Warren (1976) is especially significant; he began by seeking to identify the special characteristics faculty members

observed in able students. Three hundred faculty members in 15 California colleges and universities were asked to provide brief descriptive statements; then another 500 in 74 schools were requested to make certain judgments about phrases used by the initial 300; finally on the basis of these condensations of thoughts, around 120 faculty members were asked in telephone interviews to describe in some detail what they would like students to learn in one of their courses. Whatever their fields, the development of analytic skills was mentioned most frequently. Here is one item designed to assess that skill:

> Studies of laboratory animals have shown that overcrowding results in nervous agitation, fighting, and failure to care for the young. The growing population density of our cities is, therefore, bound to lead to increased crime and other deviant behavior.
>
> Which of the following describes the fallacy in the argument above?
>
> a. It uses ambiguous terms that shift in their meaning through the course of the argument.
>
> b. It assumes that because two things are alike in some ways, they are alike in all ways.
>
> c. It uses evidence that is not relevant to the argument.
>
> d. It assumes that the first step in a possible series of steps inevitably leads to the others.

The most compelling line of evidence supporting the proposition that classroom tests influence learning is that to which all instructors have been exposed, nevertheless this evidence goes unheeded wholesale. Faculty answers and student reactions to the two most frequently asked questions on all campuses confirm the research conclusions: "Will that be on the final?" and "Will the test be objective or essay?" If the answer to the first is "No," studying and learning cease. The answer to the second directs how students study and how much psychic energy they expend (Milton, 1982). Ask a few!

I have focused upon multiple-choice items because they are the ones used most frequently, particularly in introductory courses. Many essay items can stand improvement, too. Here is one of my favorites; it was one of five on a 50-minute exam:

> Since we know how bad inflation is, why is it that we are not now, and have not since 1972, brought inflation rates down to acceptable levels?

It is no wonder that essay items enjoy such a bad name. I suggest that at least a few good quality ones be included in test manuals; presently not many manuals have them.

I know the problem you are now thinking about: Writing test items is very time-consuming; where do I get the time? 1) begin by realizing you do not have to go it alone. Insist that publishers supply better quality items in their banks and manuals. Criticize the ones they now supply. 2) Swap items with both local colleagues and those in other schools--editing items is less time-consuming than starting from zero. For additional excellent ideas about quality items, see the work of Miller, Williams, and Haladyna (1978).

3) And finally, talk less! Meet class only twice per week and use the 50 minutes saved to prepare test items.

References

Balch, J. The influence of the evaluating instrument on student learning. American Educational Research Journal, 1964, 1, 169-182.

Cree, M. Telephone conversation. Center for Instructional Services, Purdue University, Lafayette, June, 1982.

Evans, E., Dodson, P., and Bailey, D. Assessment of instructors perceptions and use of textbook test-item manuals for measuring student achievement. Teaching of Psychology, 1981, 8, 88-90.

Glaser, R. and Bond, L. Testing: Concepts, policy, practice, and research. American Psychologist, 1981, 36, 997-1189.

Halpin, G. and Halpin, G. Testing: A key to high student achievement but low student ratings? Unpublished manuscript, Auburn University, 1980.

McMillan, J. Perceptions versus potential of multiple-choice tests. In R. Franz, R. Hopkins, and A. Toma (Eds.), Proceedings: Southern Marketing Association 1979 Conference. Lafayette: University Southwestern Louisiana Press, 1979. Pp. 70-73.

Mentzer, T. Response biases in multiple-choice test item files. Educational and Psychological Measurement, 1982, 42, 437-443.

Miller, H., Williams, R., and Haladyna, T. Beyond Facts: Objective Ways to Measure Thinking. Englewood Cliffs, N. J.: Educational Technology Publications, 1978.

Milton, O. Will That Be On the Final? Springfield, ILL: Charles C. Thomas, 1982.

Milton, O., and Eison, J. Textbook Tests: Guidelines for Item Writing. New York: Harper and Row, in press.

Semb, G., and Spencer, R. Beyond the level of recall: An analysis of higher-order educational tasks. In L. Fraley, and E. Vargas (Eds.), Proceedings of the Third National Conference on Behavior Research and Technology in Higher Education. Society for Behavioral Technology and Engineering, 1976. (Copies may be obtained from James J. Johnston, Department of Psychology, University of Florida, Gainesville, 32611.)

Warren, J. Describing college graduates in 87 phrases or less. Findings, 1976, 3, 5-8.

Wingard, J., and Williamson, J. Grades as predictors of physicians career performance: An evaluative literature review. Journal of Medical Education, 1973, 48, 311-322.